Global Positioning System and Relativity

Ghanashyam Date

The Institute of Mathematical Sciences, Chennai

http://www.imsc.res.in

shyam@imsc.res.in

May 31, 2007.
Revised on February 16, 2009.

Navigation and Positioning:

Let us begin with a flat earth. We are supposed to go to a particular destination and have a map which gives coordinates of a few easy to identify landmarks. We have lost our way. If we can locate ourselves on the map, we have a chance to take a suitable route to reach our destination. How do we locate ourselves on the map?

There are various ways to do so depending upon what further information we have about the landmarks, what tools we have at our disposals etc. Let us get primitive and further assume that landmarks are reachable by walking straight to them in a reasonable time. We then realize that we can determine the distances between the landmarks and our current position. Having gone to high school, we notice that if we can determine the distances to *at least three* landmarks (not in one line), then we can solve the three equations (for x, y):

$$(x - X_i)^2 + (y - Y_i)^2 = R_i^2 \quad , \quad i = 1, 2, 3 \ .$$

The (X_i, Y_i) are the Cartesian coordinates of the landmarks on the map. We determine R_i eg by walking to the landmarks and determine our coordinate (x, y). Now we can chart our course and reach our destination hopefully without getting lost again.

This simple scenario is in a nut-shell the problem of navigation. Several assumptions have gone into the above particular method.

- Location or positioning refers to a determining certain coordinates on a certain predetermined map;
- There is an adequate number of easily identifiable and accessible landmark objects whose map positions are known;
- We have a method of determining distances from any of the accessible landmarks.

When we get more sophisticated, our needs increase eg we may need to position ourselves in wilderness or in oceans or in sky where no clear “landmarks” are available whenever we may need them. We may also need to determine our position to a high degree of accuracy (eg for launching missiles!).

One way would be to place the “landmarks” in space i.e. satellites, so that *at least four and preferably five* are accessible any time from anywhere on earth or in near earth space. Now there are two problems: (a) distance determination is bit more sophisticated. Well, astronomers have been doing this for years. Use light and determine the light travel time to find the distance. (b) location of landmarks themselves is also non-trivial. And of course we have to choose a suitable coordinate system to refer to the various locations.

Determination of light travel time:

Let us take one problem at a time. Assume that we have a satellite in space which can transmit signals and we have a suitable receiver. We want to determine the distance between the satellite and the receiver. Due to finiteness of the speed of light, any signal pattern will be received a bit later by the receiver. Imagine now that the satellite keeps transmitting a predefined signal pattern over and over again and the receiver is also capable of generating the same signal pattern. Suppose both the satellite and the receiver begin generating the signal pattern *simultaneously*, then the pattern received by the receiver will be shifted relative to the pattern generated by the receiver with a shift related to the light travel time. It is easy for the receiver to determine the shift and hence infer the light travel time. Problem solved! But how does the receiver know *when* to begin generating the pattern (eg it may have been switched-off) without which the pattern-shift has no meaning? The receiver's clock must first synchronize with the satellite clock!

This can be done by using patterns from *five* satellites and inferring *a common instant* at which the pattern generation at the receiver should have begun. Note that four satellites is not enough since *any constant shift* will still produce *a consistent* but erroneous best fit position. A fifth satellite is needed to resolve the ambiguity of a constant shift. This can be easily seen from the two dimensional example where a fourth landmark will be needed to resolve the ambiguity of constant shift in the distances.

Thus a receiver first synchronizes its clock with that of the satellite and begins generating the predefined pattern. It compares the received pattern, finds the shift (in time) needed to infer the light travel time and then computes the distance. This determination of time shift, done by the receiver clock, will have no relation to the travel time if the two clocks are *not synchronized.* And synchronization means matching at one instant *and* having the same rate of ticking. Any errors in this will translate into an error in the distance determination.

One can obviously repeat this for four or more satellites. If all the satellites are properly synchronized, then matching with one will match with others as well. As the satellites are all "moving", we need the distance determination done at the *same* instant. This again emphasises the need for synchronization. Clearly, the satellites can also transmit their map positions so that a program at the receiver can compute the receiver's map position.

We have constructed a procedure by which positioning can be achieved, But we have also introduced further features which are subject to laws of nature eg light propagation from a satellite to a receiver and physical clocks needed in synchronization. We still have to address the issue of a map or a coordinate system.

Coordinate system and metric:

The basic idea of a coordinate system is to assign a set of numbers to a set of points such that each point has a unique assignment and each set of numbers is assigned to one and only one point. As an example consider the following procedure of assigning a pair of numbers to points on a plane. Choose a reference point (origin); choose a pair of orthogonal directions (axes); using a measuring stick, go x units along one of the axes and then y units along the other. Assign the pair (x, y) to the point so reached. Notice that if one traverses in the opposite order, one still reaches the same point so that the procedure is independent of order of traversals. But this would *not* work on the surface of a two dimensional sphere – the two different traversals reach two different points and hence the same coordinates would be assigned to two different points. This illustrates that the "most natural" procedure does not work on all types of surfaces. Another alternative which will work for all surfaces would be to draw a grid on a "sheet" and paste it on the surface without wrinkles/tearing. This would ensure the one-to-one property. However now (a) grid drawing is arbitrary (b) the pasting is also arbitrary and (c) relation between coordinate differences and physical lengths is obscured.

To appreciate the last point, consider the usual polar coordinates on the plane. This can be set up as: choose an origin; choose a reference direction; choose another direction making an angle θ with the reference direction; traverse r units along this direction and assign coordinate (r, θ) to the point so reached. Evidently, for points along a radial direction, the coordinate difference is the measured distance between the points. Along the circular arcs though the measured length would be $r\theta$. One summarizes these facts by specifying a *rule*:

$$(\Delta L)^2 \ := \ (\Delta r)^2 + r^2(\Delta\theta)^2 \ := \ g_{ij}(r,\theta)\Delta x^i \Delta x^j$$

The coefficients g_{ij} are known as the metric coefficients. In the above equation they have the form: $g_{rr}(r,\theta) = 1,\ g_{\theta\theta}(r,\theta) = r^2,\ g_{r\theta} = g_{\theta r} = 0$. For the usual Cartesian coordinate system, $g_{ij} = \delta_{ij}$. The generalization to three dimensions is immediate and obvious.

The metric thus specifies the relation between small coordinate differences and the physical length that would be measured by measuring sticks. The coordinates have a meaning only in conjunction with the metric coefficients. Furthermore, if we choose another coordinate system, we must choose corresponding metric such that the physical distance between two nearby points remains the same. Clearly, the choice of coordinates and metric coefficients must be such as to reflect the properties of the surface (or region) for which the map is being constructed.

But in our positioning problem, we are going to need assignments of time coordinates as well. So we need a four dimensional generalization. Here, the physical properties of light (or electromagnetic waves) come into play. Firstly, it has a *finite* speed of propagation and more importantly its speed (in vacuum) has a universal value independent of motion of its source or of its observer. Einstein extrapolates this observed property to be a fact of nature for all observers (first for "inertial observers" in Special Relativity and then to all observers in the "General Relativity".) This means that if we assign coordinates (t_1, X_1^i) and (t_2, X_2^i) to two events of emission of light and its subsequent absorption, then we must have $c^2(t_2 - t_1)^2 = g_{ij}(X_2^i - X_1^i)(X_2^j - X_1^j)$. Furthermore, any other assignment corresponding to relative uniform motion, the same relation must hold. Infinitesimal form of this relation is:

$$(\Delta s)^2 \ = \ g_{\mu\nu}(t, x^i)\Delta x^\mu \Delta x^\nu \ , \ \mu,\ \nu \ = 0, 1, 2, 3$$

Let us consider a couple of examples to arrive at an *interpretation* of such a "map".

Example 1: Let the map be described by coordinates (T, X^i) and a metric $g_{00} = 1, g_{ij} = -\delta_{ij},\ i = 1, 2, 3$. Consider two events (i.e. four dimensional points) defined by emission of two pulses, at two consecutive 'ticks' of a physical clock. Let the clock's motion on our map be described by $X^i(T)$ so that $V^i := \Delta X^i/\Delta T$. Let us denote, $V^2 := \delta_{ij}V^iV^j$. Then we have,

$$(\Delta\tau)^2 := \frac{(\Delta s)^2}{c^2} = (\Delta T)^2\left\{1 - V^2/c^2\right\}$$

If $V^i = 0$, i.e. the clock is at rest on our map, then the time coordinate difference ΔT is just the invariant quantity $\Delta s/c$. Thus, we interpret the $\Delta\tau$ as the difference in the readings on a physical clock *at rest* on our map. Equivalently, we can think of T as being the reading on a clock located at X^i and at rest relative to the an origin. (This can be achieved by distributing clocks all over the space and all of which are *synchronized* (i.e. show the same reading) to a clock at rest.)

Now consider another clock moving with trajectory $X^i(T)$. The two events are now defined by two consecutive ticks of a *moving* clock. Let the invariant interval be denoted by $(\Delta\tau)_{\text{moving}}$ and the corresponding time coordinate difference (which is same as the readings on the clock at rest) by ΔT_{rest}. These two are related as,

$$(\Delta\tau)_{\text{moving}} = \Delta T_{\text{rest}}\sqrt{1 - \frac{V^2}{c^2}} < \Delta T_{\text{rest}} .$$

The above map is called an *inertial coordinate system.* Its time coordinate denotes the readings of a physical clock at rest relative to the reference point of the system. Furthermore, the readings on an identically constructed moving clock progress slowly compared to the clock at rest i.e. moving clocks slow down. This is the *Special relativistic time dilation* [1]. This map describes the space-time of Special Theory of Relativity and is also known as the Minkowski space-time.

Example 2: Let the map be described by coordinates (t, r, θ, ϕ) with a metric,

$$(\Delta s)^2 = \left(1 - \frac{2GM}{c^2r}\right)(c\Delta t)^2 - \left(1 + \frac{2GM}{c^2r}\right)(\Delta r)^2 - r^2\left\{(\Delta\theta)^2 + \sin^2\theta(\Delta\phi)^2\right\}$$

Once again consider a physical clock emitting pulses at consecutive 'ticks' of the clock. This clock can be in motion, described by $(r(t), \theta(t), \phi(t))$ on the map or it can be at rest on the

[1] Rossi and Hall, in 1941, measured the flux of muons – particles similar to electron but about 200 times heavier – at the top of hill and at the bottom. The muons have a half life of about 2.2 micro-seconds when they are at rest. If there were no time dilation effect, one should have seen almost no flux of muon which was contrary to the observations. The special relativistic time dilation formula exactly gave the observed flux provided muons were travelling at about 0.98 times the speed of light. The effect has been verified at greater precision many times since.

map. For a clock at rest *and* located at very large value of r (ideally infinity), the coordinate differences coincide with the ticks of the clock just as in the previous case. Thus we conclude that *the time coordinate, t, represents the readings on a physical clock which is at rest on the map* and *located at very large values of* r [2]. By contrast, the ticks of a clock, also at rest but located at a value $r = R$, will satisfy,

$$(\Delta\tau)_{\mathrm{R}} \;=\; \Delta t_{\infty}\sqrt{1 - \frac{2GM}{c^2 R}} \;\;<\;\; \Delta t_{\infty}\ .$$

Thus, clocks at smaller values of r will run slower than clocks at larger values of r since $\Delta\tau_R < \Delta\tau_{R'}$ if $R' > R$. This is the *General Relativistic or Gravitational time dilation.* This was verified in the Pound-Rebka experiment in 1959 [3]. This map describes the space-time around a non-rotating, spherical body and is essentially the famous *Schwarzschild solution* of Einstein field equation. The M in the expression denotes the *mass* of the spherical body and r is larger than the physical radius of the body.

The two examples reveal that the rates of identically constructed clocks vary, depending upon their motion and the presence of gravitating bodies in their vicinity. For pairs of event along the trajectory of a physical (material) object ("time-like curves"), the invariant interval $\Delta\tau$ refers to the readings on some suitable physical clock. This is analogous to the statement that $(\Delta L)^2$ in the three dimensional example, is the physically measured length.

To summarise:

The setting up of a map involves a choice of coordinates, a choice of metric coefficients and some understanding of the intrinsic properties of the space-time. The crucial revelation of the general theory of relativity is that space and time must be considered together as

[2] Incidentally, the corresponding Δt intervals will also be the time intervals measured by a *locally inertial* or 'freely falling observer'.

[3] Imagine a photon being emitted in a certain atomic transition at a certain altitude. Its period can be taken as defining a clock at that altitude. An identical atomic transition at a lower altitude will similarly define another clock. Their rates, and hence the frequencies of the the emitted photons, will be related by the general relativistic time dilation expression. Thus, frequency of a photon emitted at higher altitude should be *smaller* than that emitted at a lower altitude. Hence, a photon emitted in a certain atomic transition by an atom at higher altitude *cannot* get absorbed by identical atom at lower altitude. The possibilities of absorption at the lower altitude are also hampered by the recoils of the atoms in the process, causing frequency shifts which are some 100,000 times more than the general relativistic ones. Using the Mossbauer effect – "recoil-less emission of photon" – seen in crystalline samples, it is possible to nullify the recoil effects almost completely. The absorption can also be influenced by moving the source up and down in a periodic manner which introduces a known *Doppler shift.* With this, it is possible to pick out only the gravitational shift and this was indeed verified using the 14 keV γ-ray of Fe^{59} with the two sets of atoms kept with a height difference of 22.5 meters.

a space-time and that its intrinsic properties are determined by distribution of material bodies (matter and energy) via the Einstein field equations. Once a model space-time or map is appropriately chosen, the clock rates under various conditions are determined. A synchronization, crucial for our procedure for determining positioning, must take into account these varying clock rates.

So, finally, what is the model space-time that is chosen for our global positioning problem?

The Coordinate system for GPS:

There are two main coordinate systems in use. There is one which is used for GPS computations i.e. for satellite orbit computations, synchronizations etc. This is the so called *Earth Centred Inertial (ECI)* system. The positioning however is needed mostly on the surface of earth which is *rotating*. The users would like the answer in terms of *latitude, longitude and height*. So one uses the so-called *Earth Centred Earth Fixed (ECEF)* system. A program does the necessary coordinate transformation [4]. The ECI is described by coordinates (t, r, θ, ϕ) with $r = 0$ denoting the centre of the earth and a metric of the form:

$$
\begin{aligned}
(\Delta s)^2 &= -\left\{1 + 2\frac{V(r,\theta) - \Phi_0}{c^2}\right\}(c\Delta t)^2 + \\
&\qquad \left\{1 - \frac{2V(r,\theta)}{c^2}\right\}\left((\Delta r)^2 + r^2(\Delta\theta)^2 + r^2\sin^2\theta\ (\Delta\phi)^2\right) \ , \\
V(r,\theta) &:= -\frac{GM}{r}\left\{1 - \frac{1}{2}J_2\left(\frac{R}{r}\right)^2\left(3\cos^2\theta - 1\right)\right\} \quad , \quad \text{where,} \\
GM &= 3.98600418 \times 10^{14}\ \text{meter}^3\ \text{sec}^{-2} \\
J_2 &= 1.0826300 \times 10^{-3}\ \ \text{(quadrupole moment coefficient)} \\
R &= 6378136.46\ \text{meters}\ \ \text{(Equatorial radius)} \\
\frac{\Phi_0}{c^2} &= -6.9693 \times 10^{-10}
\end{aligned}
$$

This frame is not "attached" to the earth and is non-rotating. This is signified by absence of a term of the for $\Delta t\Delta x^i$. This space-time is an approximate solution of the Einstein equation in the vicinity of the earth.

This is very similar to the metric in the example 2 above, except for two modifications. The first one involving the form of $V(r, \theta)$ incorporates the fact that earth is not exactly a sphere, it has a bulge at the equator. This is coded by the quadrupole moment of the mass distribution. There are also higher multipole moments which are too small and are ignored.

[4] The fact that ECEF is rotating requires to incorporate corrections due to the *Sagnac Effect*. This is handled at the receiver level.

The second modification is the presence of Φ_0 constant. Even for very far away from the earth, the metric coefficient g_{tt} differs from 1. Hence the coordinate time t does *not* denote the readings on a clock at infinity. Instead, it represents the readings of clock at a particular r_* determined by $V(r_*, \theta = \pi/2) = \Phi_0$. This might as well be. Our practical unit of time – the SI second – is defined by an atomic clock located at the mean sea level and at rest on the surface of earth and all atomic clocks on the satellites are synchronized with the reference atomic clock at the US Naval Observatory. But earth rotates relative to the ECI and so does the standard atomic clock. So one should use the ECEF coordinate system and find the locations where $g_{tt}^{ECEF}(r, \theta, \omega)$ equals a constant. At the equator one gets a particular value which is denoted by Φ_0. This particular equipotential surface is called the *geoid*. Thus the time now refers properly to the readings of the reference atomic clock on the geoid.

The model space-time now incorporates the matter distribution as well as refers to readings on practical clocks in sync with the SI clock. Our map is now fixed.

Determination of map positions of satellites:

As noted earlier, we need map positions of landmarks before we can determine our map positions. This means determining the locations of satellites in their orbits. The same Einstein's theory says that the satellites orbits must be (time-like) geodesics of the metric. In practice, it may suffice to use the good old Newtonian mechanics to determine the orbit equation. Observations and tracking of the satellites (telemetry) then allows to make small corrections to maintain the ideal orbits. (When the deviations are too large, the satellite broadcasts the message of its "ill-health" and the GPS receivers do not use the satellite.) This information gathered from ground by telemetry is uploaded to the satellite, roughly every two hours, and is broadcasted by the satellites. Thus, our landmarks keep announcing their identity as well as their map positions.

Synchronizing the Satellite Clocks:

In order to be able to infer the light travel time by code shifting, it is necessary for the receiver's clock to synchronize with the satellite clock. Satellites have atomic clocks, receivers don't. Since the satellites are farther away from the surface and moving, their rates suffer both the relativistic time dilations [5] and thus will *not* remain in sync with the reference atomic clock. But of course, knowing the metric, one knows the relations between these

[5] Recall that a moving clock slows down while a clock farther away from a gravitating body speeds up. A satellite at greater height will thus have a net speed-up while one at smaller height will have a net slow-down. At about 9545 km radius circular orbit the two effects cancel.

rates. It turns out that the orbit heights are such that the atomic clocks in the satellites beat faster. The clock frequency used in transmissions is 10.23 MHz. So one deliberately reduces the frequency at the launch to 10.229 999 995 43 MHz so that in orbit it reaches the desired frequency.

A story goes that when the first cesium atomic clock was put in space on 23rd June 1974, some people had doubted the general relativistic time dilation and had not incorporated the corresponding shift in the frequency of the clock. However as a safety measure, a frequency synthesizer was kept on board to make amends if the need arose. After about 20 days, indeed the clocks wend out of sync by an amount predicted by the general relativistic time dilation ($+442.5 \times 10^{-12}$ observed against $+446.5 \times 10^{-12}$ predicted) and so this correction has been incorporated since.

Now we have covered all the essential ingredients needed for a satellite based global positioning system – we have the map, the map positions of the landmarks and the workable procedure to infer light travel time to infer the distances to the landmarks. Here are now some of the actual details of the GPS used in practice.

The NAVSTAR GPS:

The name refers to the only currently deployed satellite based system and is an acronym for **NAV**igation **S**atellite **T**iming **A**nd **R**anging **G**lobal **P**ositioning **S**ystem.

It was built by the US Department of Defence during the eighties and the nineties at a cost of about 12 billion dollars. Currently it is maintained by the US Air Force at an annual cost of about 750 million dollars. It is considered a “public good” in the sense that it is non-deniable and non-rivalrous. It was put in place to be able to launch various ICBMs to destroy enemy silos. During the 90’s, non-US military users could get a less accurate data so that the GPS accuracy was limited to about 30 meters. Since 2000, this distinction has been removed.

The system consists of 24 satellites arranged in 6 groups of 4 each. Satellites in a given group lie on a plane making an angle of 55 degrees with the equator and the 4 satellites in each plane are roughly 90 degrees apart. All satellites are in nearly circular orbits (eccentricity less than 0.01) at a radius of 26,600 km with a period of 11 hours 58 minutes (passes over the same earth location once a day). At any time, from any where, at least 6 satellites are in the line of sight. Currently, there are 30 such satellites deployed, which are used by receivers for improving accuracy. These also make the system more reliable against multiple failures.

Each satellite transmits (radio waves) three different types of data in the primary *navigation message*. The *almanac* which is a coarse time info along with the satellite status; the *ephemeris* which is the orbital info which the receivers use to compute the satellite position and the clock info in two forms – Coarse/Acquisition Code or C/A and the precise or P Code. The message is 37,500 bits long. The C/A code is 1023 bits long pseudo-random code, broadcast at 1.023 MHz, repeating every milli-second. This is used to identify the satellite and subsequently to determine the travel time.

The basic limitation of accuracy comes from time delay measurements and are about 10 nanoseconds from the C/A code (and 1 nanosecond from the P code). This translates to about 3 meters (30 centimeters). There are other sources of degrading effects such as Ionospheric delays (5 meters), Ephemeris errors (2.5 meters), satellite clock errors (2 meters), multipath distortions (1 meter) tropospheric effects (.5 meters), numerical errors (1 meter).

Applications of GPS are based on the features of absolute positioning, relative movements and time transfer (synchronization). Among the applications are tracking of goods, rescue operations, power fault managements, precise mapping, detection of earth movements etc.

Some of the precursors of NAVSTAR are: Land based *LORAN* (40's), satellite based *Transit* (1960), ground based *Omega* (1970) – the first radio based, world-wide system. The first GPS satellite was launched in 1978.

References:

1. Wikipedia article at wikipedia.com ;
   Several other sources from google.com for further specilized info pertaining to actual GPS systems;

2. A precis of General Relativity as used in GPS is nicely given in an article by Charles Misner, arxiv:gr-qc/9508043;

3. The review article by Neil Ashby available at www.livingreviews.org/lrr-2003-1, contains more details discussions.